Effective Surgical Visual Question Answering Without Ground-Truth Video Descriptions

Team name: **Capybara**

Author: **Quan Huu Cap**

Affiliation: **Aillis, Inc., Tokyo, Japan**

Link to code repo: https://github.com/huuquan1994/surgvu25-cat2-submission

Private dataset used (e.g. private videos, additional annotations of challenge training set, etc): No

Introduction

Vision-Language Models (VLMs) can describe surgical videos at general level but often fail to identify surgical tools and organs accurately. For this task, a seemingly straightforward solution is to fine-tune VLMs with the provided videos and captions. However, the dataset is noisy and contains only 21 unique captions (based on "match_description"), which is insufficient for training. Preliminary fine-tuning experiments with the provided descriptions also yielded unsatisfactory results.

In the public sample set (11 videos), most questions focus on surgical tools, organs, and actions (e.g., cutting tissues, using forceps). Assumably, if tools and organs are detected, VLMs could infer video actions without relying on ground-truth descriptions (e.g., if scissors appear, likely tissues are being cut). In this work, we propose a method to detect surgical tools and organs in video frames and combine this with general video descriptions for surgical visual question answering (VQA).

Methodology

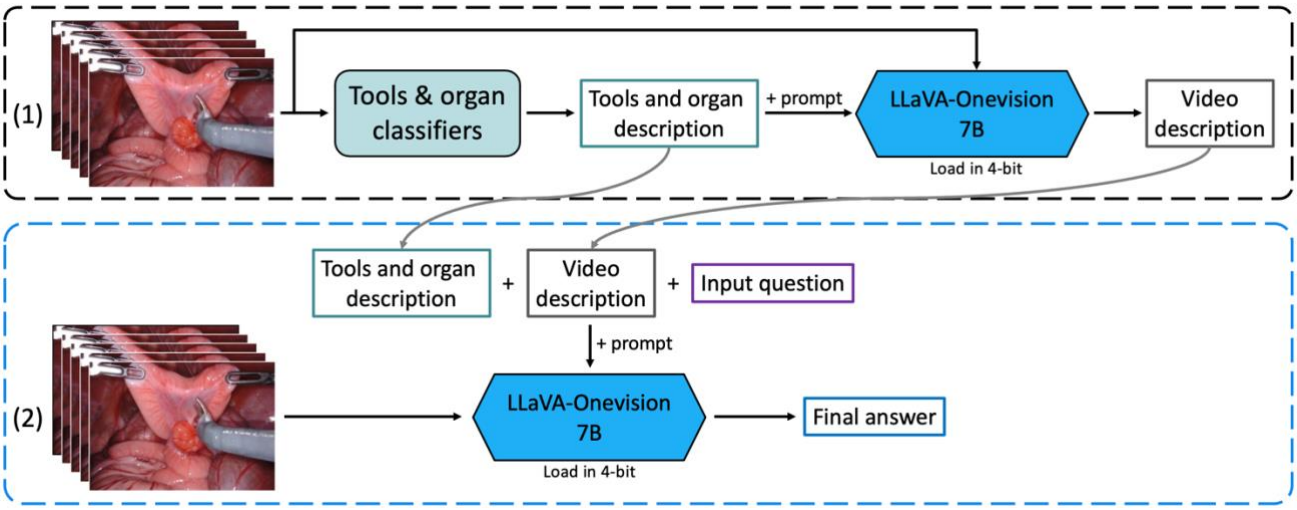

Fig.1. The overview of the proposed effective surgical VQA method.

Fig. 1 illustrates the overview of our proposed method. Our method consists of two steps: (1) create a video description; and (2) get the final answer. *First*, surgical tools and organ classifiers are employed to classify

the tools and organ present in the video frames, obtaining the tools and organ description. Then, this tools and organ description together with a prompt is used to generate the video description. *Second*, the tools and organ description, the generated video description, the input question, and the video frames are used to generate the final answer. Here, the VLM and video frames are the same at both steps. We chose to use the LLaVA-OneVision 7B model [1] based on our preliminary experiments. This is the original model without any additional fine-tuning. More on the VLM selection will be discussed later.

For each frame, the black margins on the left and right were cropped, and the tool list area at the bottom was obscured using Gaussian blur before inference. For more details on video pre-processing, please refer to our GitHub repository.

The Surgical Tool Classifier

Because the tool presence labels in the provided dataset are noisy and the clip lengths vary widely (0–5000 seconds), creating a reliable surgical tool classification dataset is very challenging. To address this, we leveraged a public dataset from the SurgToolLoc Challenge (MICCAI 2022) [2], which contains ~24,600 clips of 30 seconds each with tool presence annotations. Although this dataset is also somewhat noisy, it substantially reduces the effort required for data preparation.

We removed clips that were too small or large, and built a multi-label dataset with 22k training clips and 2.4k validation clips (30 frames each). Video frames were pre-processed same way as described above. We trained EfficientNetV2-Small [3] with 512×512 input, applying label smoothing to handle noisy labels, and achieved a 97% macro F1-score. At test time, surgical tools are detected frame by frame.

The Organ Classifier

There are seven types of organs from the challenge dataset description (`match_description` in CSV files). With the associate videos, we extracted 2700 clips for training the organ classifier (2300 clips for training, 400 clips for validation; 30 frames/clip). Assuming each clip has one organ, this is the single-class classification. The model, input size, and pre-processing method are the same as in the surgical tool classifier. We achieved a macro F1-score of 98%. At test time, the organ prediction is based on voting (e.g., if rectum is predicted in most frames, the final prediction is rectum).

Generating the Video Description and Final Answer

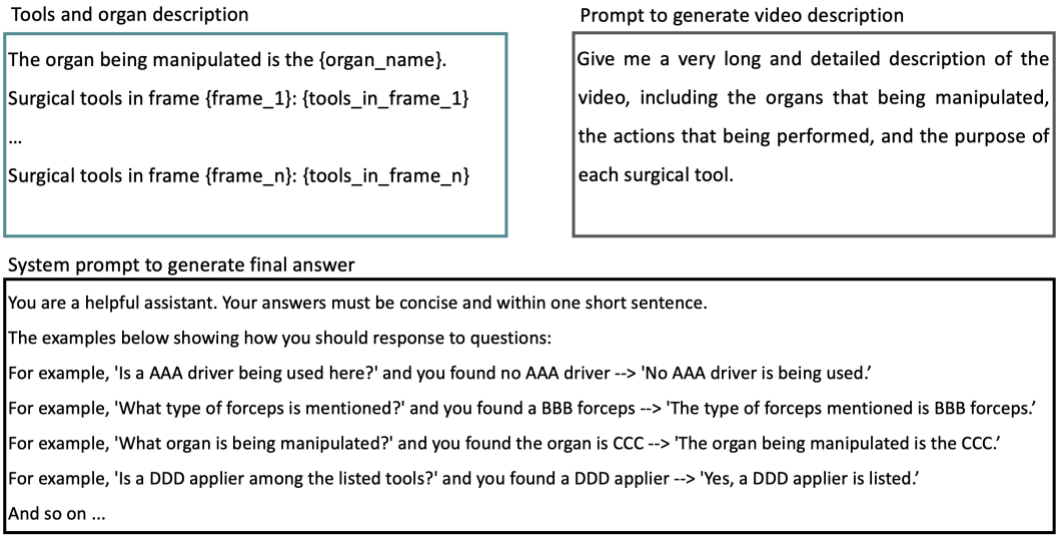

Fig.2. The template of tools and organ description, and the prompts to generate video description and final answer.

Fig. 2 shows the template of the tools and organ description, and the prompts for generating video description and final answer.

Because the sample ground-truth answers are provided in short-sentence form, we design the system prompt to generate short answers. The prompt was selected by iteratively testing different candidates and choosing the one that maximized the BLEU score on the public sample set (11 videos). It is worth noting that some answers in the public sample are wrong and need correction. We provide the corrected version used for validation in our GitHub repository.

Results & Discussion

Our best final submission samples 5 frames/video as input. We achieved 0.4237 BLEU in the Prelim phase, and 0.4215 BLEU in the Final phase (rank 1st). Among other final submissions, the variant with 21 frames/video input got 0.4022 BLEU (rank 2nd), while replacing LLaVA-OneVision with the reasoning model R-4B [4] yielded lower result with 0.3448 BLEU (rank 4th).

Throughout the challenge, we experimented with several VLMs and observed that models using the Qwen text encoder (e.g., Qwen2, Qwen3) were easier to control the output compared to those based on LLaMA or BERT. We also found that VLMs with reasoning capabilities better captured video semantics, but were more difficult to control. Nevertheless, with appropriate configuration, such models could be a promising approach.

Additionally, we observed that VLMs quantized to 4-bit precision showed inconsistent behavior across GPU architectures (e.g., Turing, Ampere, Hopper), producing different outputs for the same input. Since the evaluation platform uses Tesla T4 (Turing), we recommend testing methods on the same architecture for reproducibility. As future work, enriching video descriptions through robust object detection or collecting higher-quality captions could further improve performance.

References

[1] B. Li *et al.*, “Llava-onevision: Easy visual task transfer,” 2024, arXiv:2408.03326

[2] A. Zia e*t al.*, “Intuitive Surgical SurgToolLoc Challenge Results: 2022-2023,” 2023, arXiv:2305.07152

[3] M. Tan and Q. V. Le, “EfficientNetV2: Smaller Models and Faster Training,” 2021, arXiv: 2104.00298

[4] J. Jiang *et al.,* “R-4B: Incentivizing General-Purpose Auto-Thinking Capability in MLLMs via Bi-Mode Annealing and Reinforce Learning” 2025, arXiv:2508.21113